Human Genome Sequence and Variation

Dr. S Hosseini-asl

The goals of the different phases of the Human Genome Project

(1) determine the linkage map of the human genome

(2) construct a physical map of the genome by means of cloning all fragments and arrange them in the correct order

(3) determine the nucleotide sequence of the genome

(4) provide an initial exploration of the variation among human genomes.

- As of October 2004 about 93% of the human genome (which corresponds to 99% of the euchromatic portion of the genome) had been sequenced to an accuracy of better than one error in 100,000 nucleotides.
- The length of the human chromosomes ranges from ~46 Mb to ~247 Mb.
- The average GC content of the human genome is 41%.
- This varies considerably among the different chromosomes and within the different bands of each chromosome. Chromosomal bands positive for Giemsa staining have lower average GC content of 37%, while in Giemsa-negative bands the average GC content is 45%. Interestingly, Giemsa-negative bands are gene rich regions of DNA.

NCBI Build 36.1, Mar. 2006 Assembly (hg18)

Chr Name	Assembled Size (inc. Gaps)	Sequenced Size	Total Gap Size	Non-Euch. Gap Size
1	247249719	224999719	22250000	20240000
2	242951149	237712649	5238500	4200000
3	199501827	194704827	4797000	4490000
4	191273063	187297063	3976000	3010000
5	180857866	177702766	3155100	3083000
6	170899992	167273992	3626000	3008000
7	158821424	154952424	3869000	3184000
8	146274826	142612826	3662000	3000000
9	140273252	120143252	20130000	18000000
10	135374737	131624737	3750000	2380000
11	134452384	131130853	3321531	3257000
12	132349534	130303534	2046000	1471000
13	114142980	95559980	18583000	17933000
14	106368585	88290585	18078000	18078000
15	100338915	81341915	18997000	18260000
16	88827254	78884754	9942500	9805000
17	78774742	77800220	974522	220000
18	76117153	74656155	1460998	1363998
19	63811651	55785651	8026000	8016000
20	62435964	59505253	2930711	1773661
21	46944323	34171998	12772325	12769767
22	49691432	34851332	14840100	14430000
X	154913754	151058754	3855000	3000000
Y	57772954	25652954	32120000	30500000
M	16571	16571	0	0
Overall Chrom	3080436051	2858034764	222401287	205472426

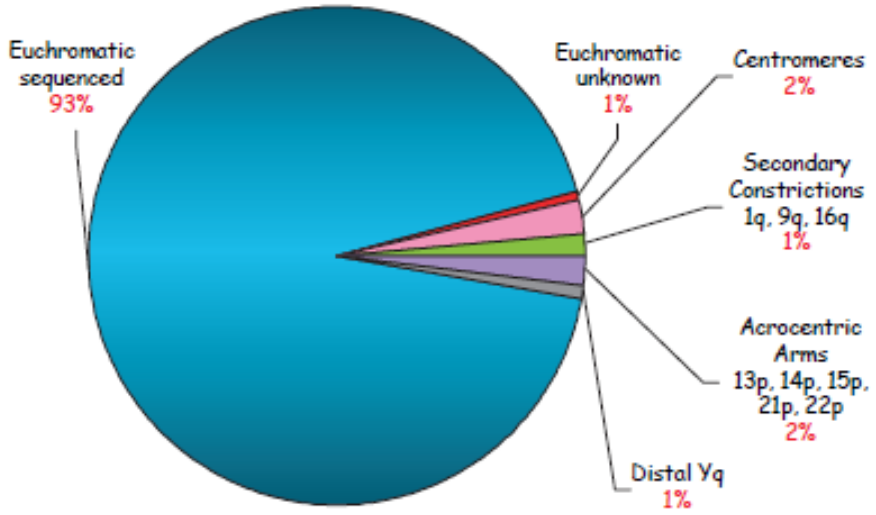

Fig. 2.1 Pie chart of the fractions of the genomes sequenced (*blue*) and not sequenced (*non-blue*)

- The sequence of the human genome is freely and publicly available on the following genome browsers:

(a) http://genome.ucsc.edu/

(b) http://www.ensembl.org/

(c) http://www.ncbi.nlm.nih.gov/genome/guide/human/

ENCODE

- There is now a considerable effort internationally to identify all the functional elements of the human genome.
- A collaborative project called ENCODE (ENcyclopedia Of DNA Elements) is currently in progress with the ambitious objective to identify all functional elements of the human genome

- The current classification of the functional elements of the genome contains:
- 1. Protein-coding genes
- 2. Noncoding, RNA-only genes
- 3. Regions of transcription regulation
- 4. Conserved elements not included in the above categories

CCDS

- The so-called CCDS set (*consensus coding sequence) is built by consensus among the European* Bioinformatics Institute (http://www.ebi.ac.uk/), the National Center for Biotechnology Information (http:// www.ncbi.nlm.nih.gov/), the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/), and the University of California, Santa Cruz (UCSC; http://www.cbse. ucsc.edu/). At the last update (5 July 2009; genome build 36.3) CCDS contains 17,052 genes. This is the minimum set of protein-coding genes included in all genomic databases. The reference sequence (RefSeq) collection of genes of the NCBI contains 20,366 protein- coding gene entries (http://www.ncbi.nlm.nih.gov/RefSeq/); the UCSC collection of genes contains 23,008 entries http://genome.ucsc.edu/); the Ensembl browser contains 21,416 entries (23 June 2009; build 36; http://www.ensembl.org/Homo_sapiens/Info/ Stats Table). The total number of annotated exons listed in the Ensembl database is 297,252 (23 June 2009; build 36). The discrepancy among the databases reflects the ongoing and unfinished annotation of the genome.

The different classes of RNA-only genes

1. rRNAs

- *Ribosomal RNA (rRNA) Genes: ~650*–900. These are genes organized in tandemly arranged clusters in the short arms of the five acrocentric chromosomes (13, 14, 15, 21, and 22). The transcripts for 28 S, 5.8 S, and 18 S rRNAs are included in one transcription unit, repeated 30–50 times per chromosome. These tandemly arranged genes are continuously subjected to concerted evolution, which results in homogeneous sequences due to unequal homologous exchanges. The transcripts for the 5 S rRNAs are also tandemly arranged, and the majority map to chromosome 1qter.

28 S (components of the large cytoplasmic ribosomal subunit)	~150–200
5.8 S (components of the large cytoplasmic ribosomal subunit)	~150–200
5 S (components of the large cytoplasmic ribosomal subunit)	~200–300
18 S (components of the small cytoplasmic ribosomal subunit)	~150–200

2. tRNAs

- *Transfer RNA (tRNA) : ~500 (49 Types).*
- *At the last* count there are 497 transfer RNA genes (usually 74–95 nucleotides long) encoded by the nucleus and transcribed by RNA polymerase III (additional tRNAs are encoded by the mitochondria genome). There are also 324 tRNA pseudogenes . The tRNA nuclear genes form 49 groups for the 61 different sense codons.
- Although the tRNA genes are dispersed throughout the genome, more than 50% of these map to either chromosomes 1 or 6; remarkably 25% of tRNAs map to a 4-Mb region of chromosome 6.

3.snRNAs

- *Small Nuclear RNA (snRNA): ~100.*
- These are heterogeneous small RNAs. A notable fraction of these are the spliceosome RNA genes many of which are uridine-rich; the U1 group contains 16 genes, while U2 contains six, U4 4, U6 44, and the other subclasses are represented by one member. Some of these genes are clustered, and there is also a large number of pseudogenes (more than 100 for the U6 class).

4. snoRNAs

- *Small Nucleolar RNA (snoRNA) : ~200.*
- *This is a large* class of RNA genes that process and modify the tRNAs and snRNAs .
- There are two main families:
- **C/D box snoRNAs** that are involved in specific methylations of other RNAs; and **H/ACA snoRNAs**, mostly involved in site-specific pseudouridylations. Initially, there were 69 recognized in the first family and 15 in the second ; however, the total number is probably larger.
- A cluster of snoRNAs maps to chromosome 15q in the Prader–Willi syndrome region (at least 80 copies); deletions of which are involved in the pathogenesis of this syndrome. Another cluster of snoRNAs maps to chromosome 14q32 (~40 copies).
- The majority of snoRNAs map to introns of protein-coding genes and can be transcribed by RNA polymerase II or III.

5. miRNAs

- *Micro RNAs (miRNA) : (706 Entries on 26 June* 2009).
- These are single-stranded RNA molecules of about 21–23 nt in length that regulate the expression of other genes.
- miRNAs are encoded by RNA genes that are transcribed from DNA but not translated into protein; instead they are processed from primary transcripts known as pri-miRNA to short stem-loop structures called pre-miRNA and fi nally to functional miRNA. Mature miRNA molecules are complementary to regions in one or more messenger RNA (mRNA) molecules, which they target for degradation. A database of the known and putative miRNAs, and their potential targets, can be found in http://microrna.sanger.ac.uk/. miRNAs have been shown to be involved in human disorders.

6. *LincRNAs*

- *Large Intervening Noncoding RNAs (LincRNAs)* : ~1,600.
- This new class has been recently identified using trimethylation of Lys4 of histone H3 as a genomic mark to observe RNA PolII transcripts at their promoter, and trimethylation of Lys36 of histone H3 marks along the length of the transcribed region to identify the spectrum of PolII transcripts. Approximately 1,600 such LincRNA transcripts have been found across four mouse cell types (embryonic stem cells, embryonic fibroblasts, lung fibroblasts, and neural precursor cells).
- Among the "exons" of these LincRNAs, approximately half are conserved in mammalian genomes, and are thus present in human. Since this class was described in 2009, further work is needed for its characterization and validation, as well as the potential overlap of its members with the other classes.

7. *Other Noncoding RNAs*

- *Other Noncoding RNAs:* ~1,500.
- The field of noncoding RNA series is constantly expanding. Some of these RNA genes include molecules with known function such as the telomerase RNA, the 7SL signal recognition particle RNA, and the XIST long transcript involved on the X-inactivation.
- There are also numerous antisense noncoding RNAs, and the current effort to annotate the genome suggests that a substantial fraction of the transcripts are noncoding RNAs.

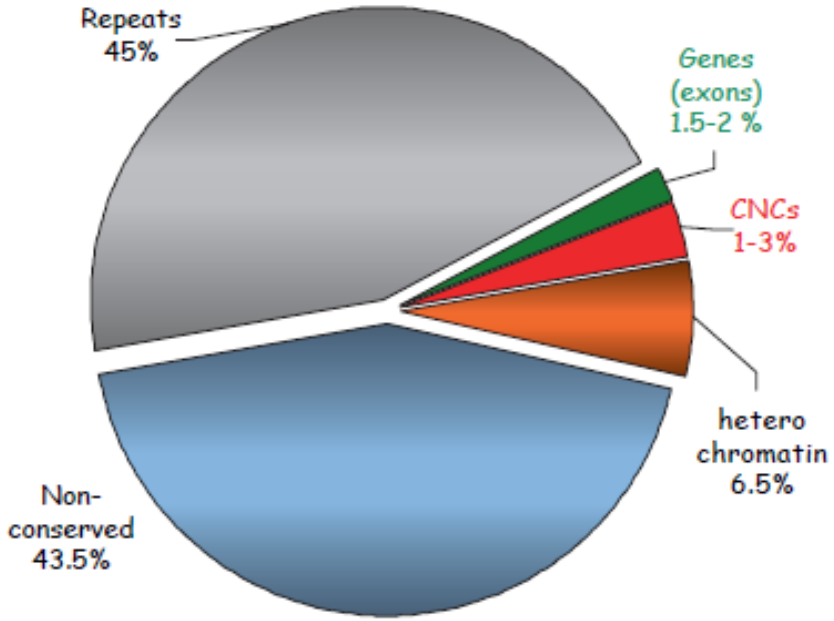

Fig. 2.7 The pie-chart depicts the different fractions of the genome. *CNCs*, conserved noncoding sequences

Mitochondrial Genome

- In human cells there is also the mitochondrial genome, which is 16,568 nucleotides long and encodes for 13 protein-coding genes, 22 tRNAs, one 23 S rRNA, and one 16 S rRNA (http://www.mitomap.org).
- The mitochondria genome-encoded genes are all essential for oxidative phosphorylation and energy generation in the cell. Each cell has hundreds of mitochondria and thousands (10^3–10^4) of mitochondria DNA (mtDNA) copies. Human mtDNA has a mutation rate ~20 times higher than nuclear DNA. The inheritance of mtDNA is exclusively maternal (the oocyte contains 10 5 mtDNA copies). Several human phenotypes are due to pathogenic mutations in the mitochondrial genome.

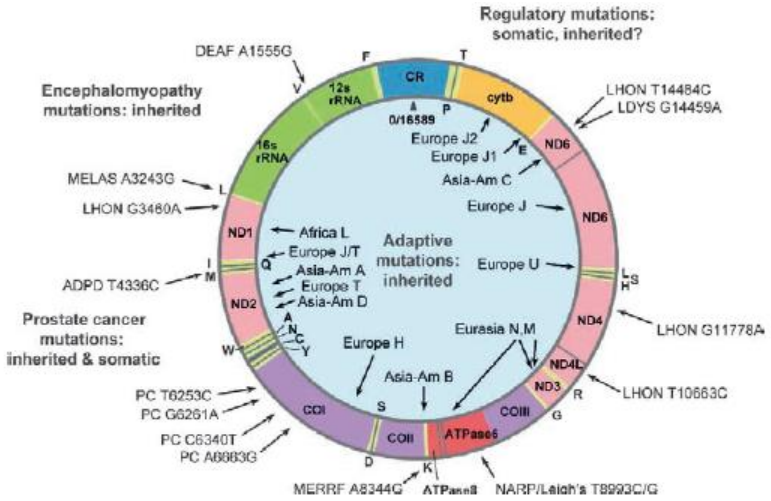

Fig. 2.14 Schematic representation of the circular mtDNA, its genes, its clinical relevant mutations, and certain polymorphic markers. *Letters within the ring* depict the genes encoded. *Letters on the outside* indicate amino acids of the tRNA genes. *CR*, the control of replication region that contains promoters for the heavy and light strands. *Arrows outside* show the location of pathogenic mutations. (From [142]

Genomic Variability

1. *Single Nucleotide Polymorphisms*

- The majority of the DNA variants are single nucleotide substitutions commonly known as SNPs (single nucleotide polymorphisms). The first SNPs were identified in 1978 in the laboratory of Y.W. Kan 3' to the β–globin Gene.

Fig. 2.15 Schematic representation of a single nucleotide polymorphism. Allele 1 has a C in the sequence, while allele 2 contains a T in the same position

- There is on average one SNP in ~1,000 nucleotides between two randomly chosen chromosomes in the population. Many of these SNPs are quite common. A common SNP is that in which **the minor allele frequency** (MAF) is more than 5%. On average two haploid genomes differ in ~3,000,000 SNPs.
- In addition, there is a large number of rare (MAF < 1%) or near-rare (MAF between 1% and 5%) SNP variants that could be identified by the genome sequencing of various individuals. The majority of heterozygous SNPs in the DNA of a given individual are relatively common in the population; on the other hand, most of the SNPs discovered in a population are more likely to be rare. The NCBI SNP database contains 25 million common and rare SNPs (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi; version 130; July 2009).
- Of those, ~301,000 are in the protein-coding regions of genes, and ~188,000 result in amino acid substitutions (nonsynonymous substitutions). An international project known as HapMap (http://www.hapmap.org/) has completed the genotyping of ~4,000,000 common SNPs in individuals of different geo-ethnic origins (4,030,774 SNPs in 140 Europeans; 3,984,356 in 60 Yoruba Africans; 4,052,423 in 45 Japanese and 45 Chinese; http://www.hapmap.org/downloads/index.html.en).

2. Short Sequence Repeats

- Short sequence repeats (SSRs) are polymorphic variations due to a different number of short sequence repeat units, first described by Wyman and White (then called VNTRs, variable number of tandem repeats

Fig. 2.18 An example of a dinucleotide SSR with three alleles in the population: the *blue* allele with $(CA)_{13}$ repeats, the *red* allele with $(CA)_{16}$, and the *green* allele with $(CA)_7$

3. *Insertion/Deletion Polymorphisms (Indels)*

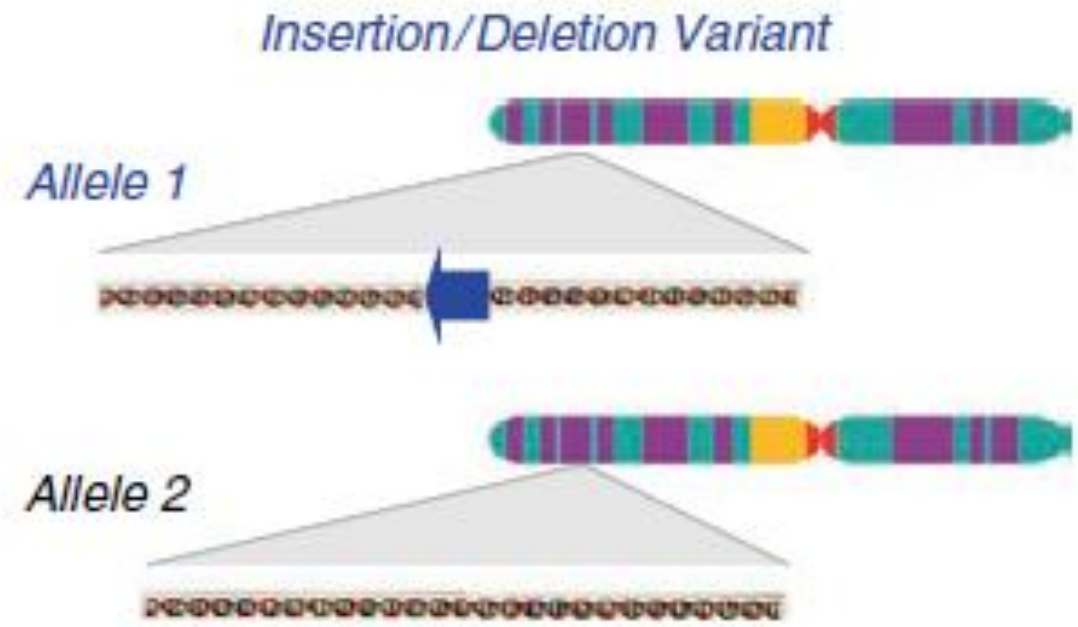

Fig. 2.19 Schematic representation of a polymorphic locus due to insertion deletion of a genomic element, shown as a *blue arrow*

4. Copy Number Variants (CNVs)

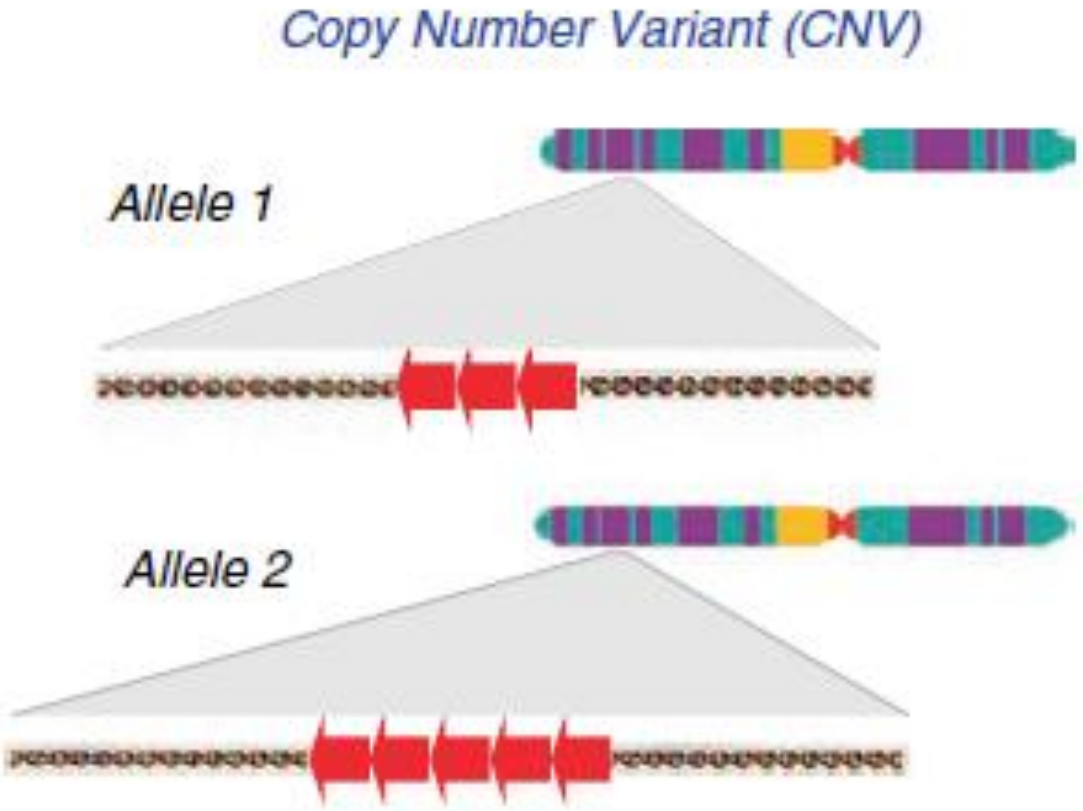

Fig. 2.20 Schematic representation of copy number variation in the human genome. For explanation, see text. Allele 1 in the population contains three copies of a sequence (*red arrowheads*), while allele 2 contains five copies

5. Inersions

- An example of a common inversion polymorphism involves a 900-kb segment of chromosome 17q21.31, which is present in 20% of European alleles but it is almost absent or very rare in other populations. These variants are difficult to identify and most of them have been detected by sequencing the ends of specific DNA fragments and comparing them with the reference sequence

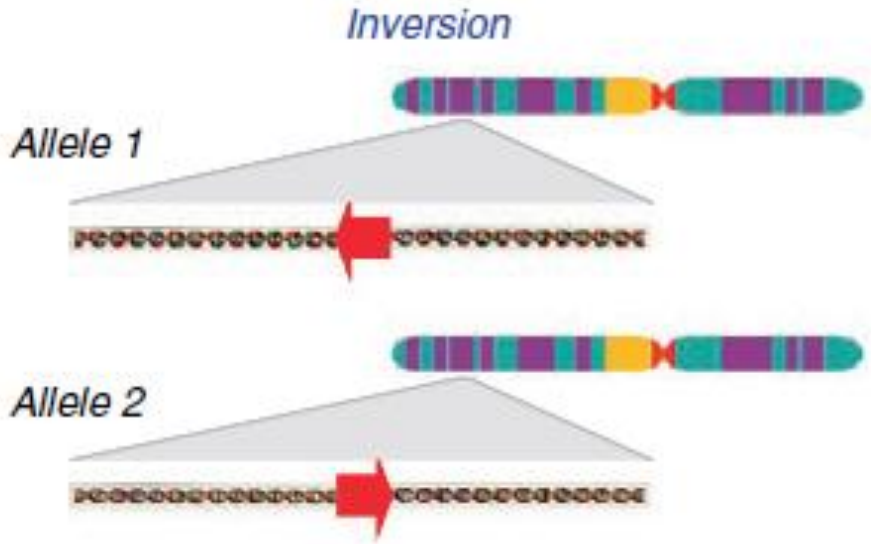

Fig. 2.21 Schematic representation of a polymorphic inversion shown as a *red arrowhead*

6. Mixed Polymorphisms

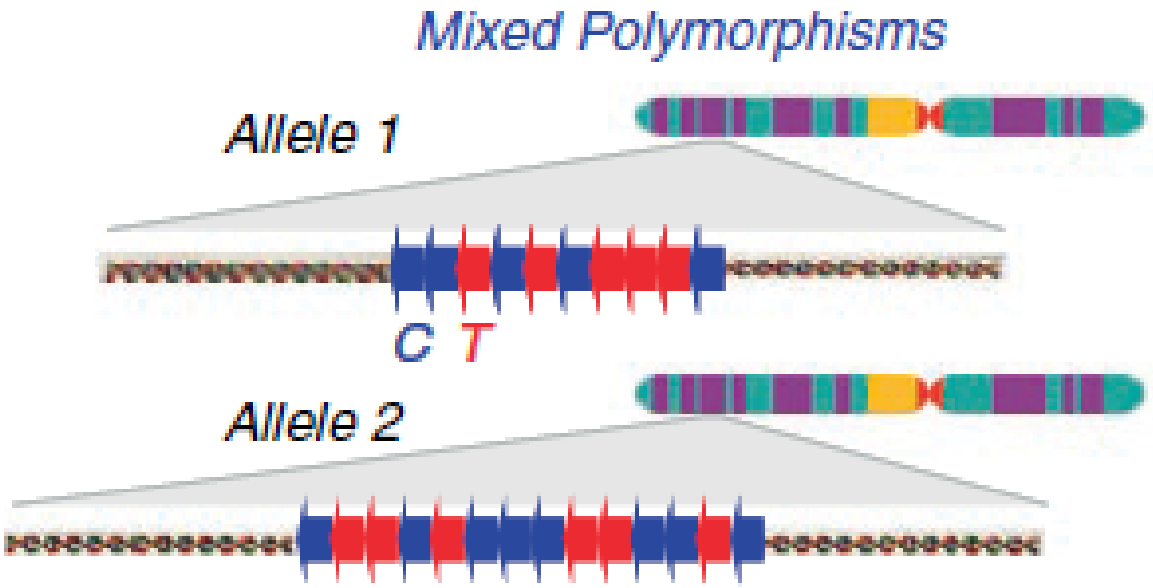

Fig. 2.22 Schematic representation of a highly polymorphic region of the genome with a mixed polymorphism that includes SNPs in the copies of CNVs or SSRs. The copies of the repeat are shown as *arrowheads*; the *blue/red* color of the repeats designates the SNP in them (*blue* for C and *red* for or T)

Genome Variation as a Laboratory Tool to Understand the Genome

- 1. Create linkage (genetic) maps of human chromosomes. This has allowed the initial mapping of the human genome and it was a prerequisite for the sequence assembly.
- 2. Map the genomic location of monogenic phenotypes to human chromosomes by linkage analysis. A large number of such phenotypes have been mapped to small genomic intervals because of the genotyping of members of affected families. Positional cloning of pathogenic mutations was subsequently possible.
- 3. Map the genomic location of polygenic phenotypes to human chromosomes by genome wide linkage and association studies.
- 4. Allow fetal diagnosis and carrier testing by linkage analysis of the cosegregation of a polymorphic marker and the phenotype of interest.
- 5. Perform paternity and forensic studies. A whole field was developed mainly with the use microsatellite SSR variants.
- 6. Study genome evolution and origin of pathogenic mutations.
- 7. Study the recombination rate and properties of the human genome.
- 8. Study the instability of the genome in tumor tissues.
- 9. Identify loss-of-heterozygosity in human tumors.
- 10. Study uniparental disomy and thus help with understanding genomic imprinting.
- 11. Study parental and meiotic origin, and decipher the mechanisms of nondisjunction.
- 12. Study population history and substructure.